Reddit Comment Extraction

Name: Tanvica Samudrala
IU email: tasamud@iu.edu
15/03/2025

Abstract

Social media platforms like Reddit give everyone a chance to discuss topics freely, but sometimes conversations become harmful or toxic. For this assignment, I built a network from Reddit comments to understand user interactions, community structures, and how toxicity influences engagement. I collected around 5,900 comments from a popular Reddit discussion using PRAW, structured this data into a network with users as nodes and replies as edges, and applied methods like Louvain community detection, clustering, and centrality analysis to explore the interactions. Additionally, I classified comments into Non-Toxic, Mildly Toxic, and Highly Toxic categories using the Detoxify model to investigate toxicity trends. My analysis shows how toxic comments and users behave differently in the network, helping us understand where negativity emerges and how it affects online conversations. The findings can support strategies for better community moderation and healthier social media interactions.

1. Introduction and Background

Social media platforms, such as Reddit, enable millions of users to engage openly on diverse topics, facilitating discussions that can range from constructive debates to aggressive interactions. While these platforms democratize conversations, they also often experience issues with toxicity, which includes negative, abusive, or disruptive behaviors. Understanding how toxicity influences interactions, shapes community structures, and impacts user engagement is crucial for improving moderation and promoting healthier digital environments.

This research explores user interactions and the dynamics of toxicity within a highly active Reddit discussion on music, particularly focusing on how toxicity affects conversation patterns, community structures, and user engagement. I utilized Python and the Reddit API (PRAW) to extract around 5,900 comments from a prominent Reddit thread, structuring the conversation as a network. In this network, nodes represent users, and directed edges represent reply interactions, forming a visual and analytical basis to study interactions.

To understand the behavioral aspect of conversations, each comment was analyzed using Detoxify, an NLP model that assigns toxicity scores to text, classifying comments as Non-Toxic, Mildly Toxic, or Highly Toxic. This allowed me to examine whether and how toxicity influences user interactions and overall community dynamics. Additionally, I employed advanced network analysis techniques including community detection (Louvain Method), hierarchical clustering (Girvan-Newman Method), and various centrality measures (Degree, Betweenness, Eigenvector) to reveal structural and behavioral insights.

By network analysis methods this research seeks to understand not only the structural aspects of online interactions but also the spread and influence of toxic behavior. Ultimately, this comprehensive approach aims to provide actionable insights for improving moderation strategies and fostering healthier digital community environments.

2. Data Collection

For this assignment, the data I collected includes user comments from a highly active Reddit discussion thread related to music, using the Python Reddit API Wrapper (PRAW). The extracted dataset consists of approximately 5,900 comments, capturing essential details such as anonymized usernames, comment content, timestamps, upvote counts, and reply interactions.

I structured this data into a network format to better understand how users interact with each other. In this network, each node is an individual Reddit user who commented, and each edge represents a direct interaction where one user replied to another. This setup allowed me to visualize the entire discussion as a clear web of interactions and enabled detailed analysis of user behavior. Additionally, each comment was analyzed using Detoxify, an NLP tool, to assign toxicity scores. This approach helped categorize users into Non-Toxic, Mildly Toxic, or Highly Toxic groups, adding an extra layer to my analysis by highlighting the behavioral aspects of interactions alongside structural connections.

Defining the Network:

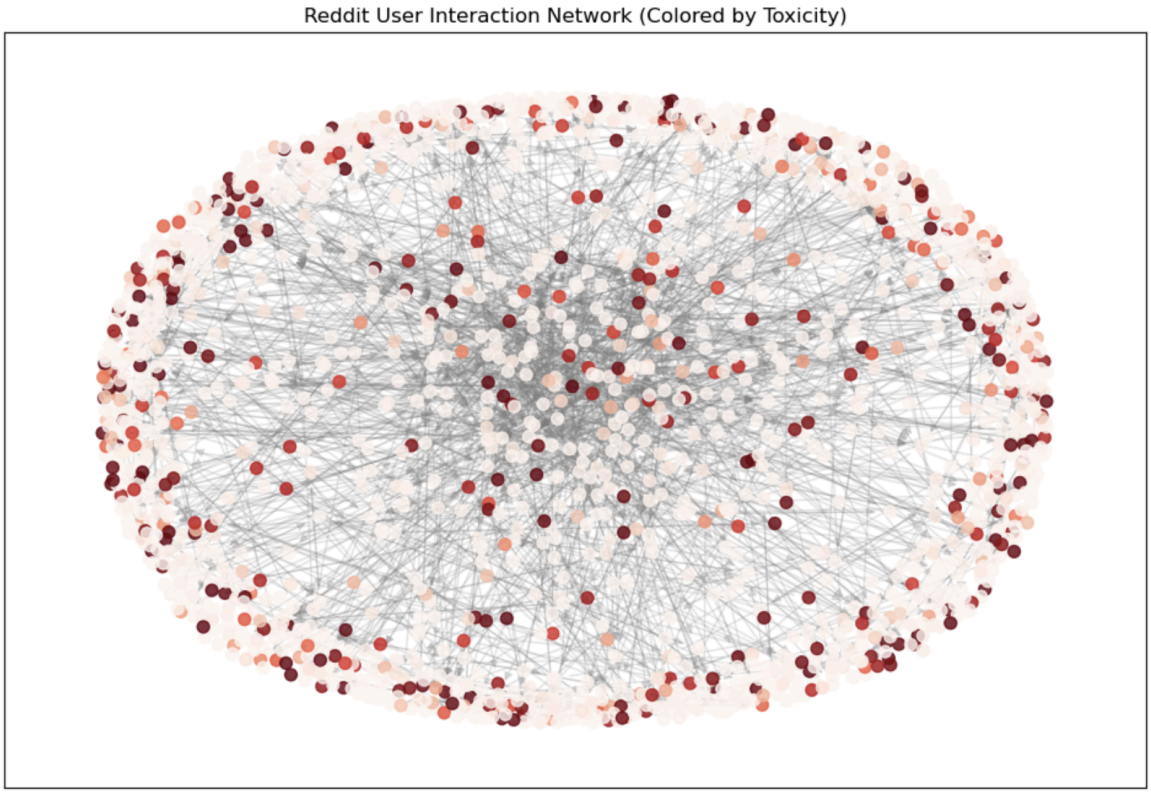
In this, the Reddit discussion is represented as a graph-based network, where users and their interactions form a structured web of connections. This network helps visualize how conversations flow, which users engage with each other, and how toxicity is distributed across different groups.

- Nodes (Users): Each unique Reddit user who participated in the discussion is represented as a node in the network. Every user who posted a comment or replied to another user is included as a node.
- Edges (Reply Interactions): A directed edge is created between two users when one user replies to another. The direction of the edge shows the flow of conversation, meaning who is responding to whom.

Networks help us visualize how people interact online by using nodes (users) and edges (replies). I used NetworkX to build the interaction network from Reddit comments, where every user is a node and every reply between two users creates an edge. This structure allows for advanced community detection, clustering, and centrality analysis, helping to uncover patterns.

3. Analysis

In this assignment, I used multiple network analysis methods to explore how Reddit users interact and how their behavior, specifically toxicity, impacts discussions. First, I visualized the network by representing each user as a node and each reply as an edge, clearly marking the nodes based on toxicity levels—dark dots represented highly toxic users, and lighter dots indicated less toxicity. This visualization showed that highly toxic users usually appeared towards the outer edges of the network rather than at its center, suggesting that toxicity is concentrated among users who are less connected with the core conversation, forming smaller isolated groups.

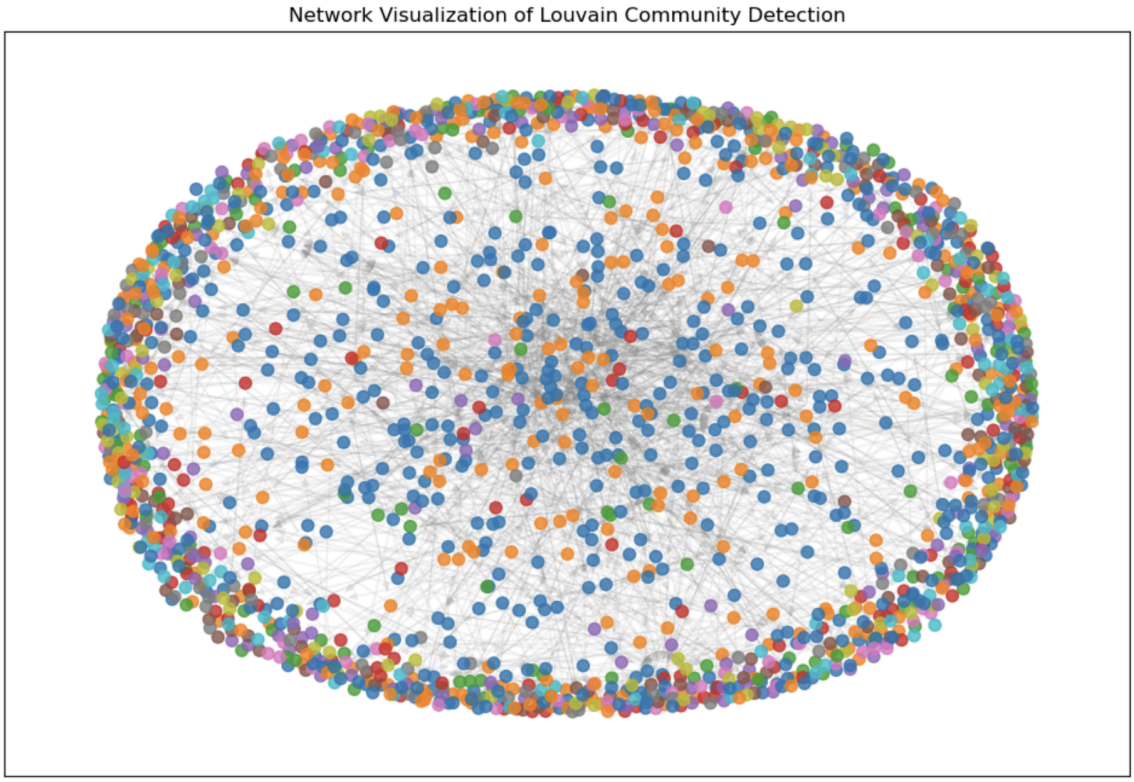


The above Reddit User Interaction Network shows how users interact and how toxicity spreads in conversations. The darker dots represent highly toxic users, while the lighter dots are less toxic or neutral. From the visualization, we can see that toxic users tend to be more on the edges, meaning they are not deeply involved in discussions. The center of the network is densely connected, and most users there seem to have lower toxicity levels. This suggests that toxic comments often come from users who are less engaged with the community. This pattern is important for platforms trying to reduce harmful content, as it shows that focusing on these outer, less-connected users might be key to controlling toxicity.

The Reddit users' analysis is based on their average toxicity score into three groups: Non-Toxic, Mildly Toxic, and Highly Toxic. The results show that a majority of users (1270) fall into the Non-Toxic category, while 330 users are Mildly Toxic, and 192 users are Highly Toxic. This

distribution indicates that most users engage in relatively civil discussions, but a notable fraction of the community exhibits toxic behavior.

By segmenting users this way, platforms can identify and monitor harmful behavior more effectively. For instance, moderation efforts could focus on the Highly Toxic group to reduce negativity, while those in the Mildly Toxic group could be encouraged toward more positive interactions. Additionally, understanding these toxicity levels helps in designing better content filtering algorithms and improving online community health.



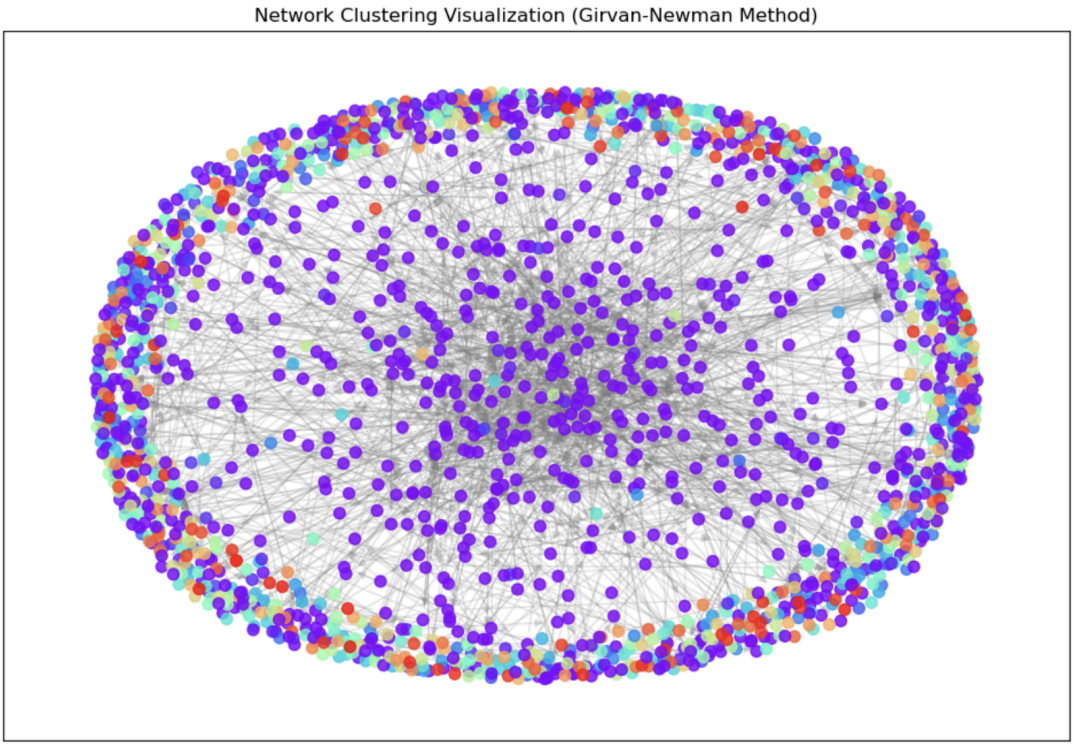
The Louvain community detection analysis revealed a very fragmented Reddit user interaction network, indicating that conversations typically happen in small, specialized groups rather than in large-scale discussions. Over 800 micro-communities were initially found, suggesting that users prefer interacting within smaller circles, aligning well with Reddit's structure of niche threads and sub-communities.

To gain deeper insights, I reclassified these numerous communities based on their toxicity levels—Non-Toxic, Mildly Toxic, and Highly Toxic—rather than solely structural interactions. This helped highlight user behaviors within these groups. For instance, some communities showed extremely high average toxicity, indicating these groups were hotbeds for negative interactions, while others remained largely positive or neutral. This combination of structural analysis (who interacts with whom) and behavioral analysis (who contributes to toxicity) offers valuable information for platforms to identify problematic interactions early and implement effective moderation strategies.

Interestingly, some communities are notably larger than others, as seen in the visualization. The distribution of community sizes suggests that a few influential groups dominate the discussion, while many smaller ones exist on the periphery. This could indicate that certain users or topics are central to the discourse, drawing more engagement than others. The presence of small, isolated communities suggests that Reddit fosters niche discussions, where users with shared interests interact in self-contained groups rather than merging into a broader conversation.

One critical insight is the potential relationship between toxicity and community structure. Since we have separately classified users into Non-Toxic, Mildly Toxic, and Highly Toxic categories, comparing these groups against the detected communities helps understand whether toxicity is concentrated in specific groups or spread throughout the platform.

From a moderation perspective, identifying such highly toxic communities is crucial. Platforms like Reddit could use automated interventions, such as increased content moderation, flagging mechanisms, or restricting certain discussions, to mitigate the spread of toxicity. On the other hand, the presence of large, non-toxic communities suggests that positive, constructive discourse is also thriving, demonstrating that Reddit is not entirely dominated by negative interactions. Understanding these patterns provides valuable insights for social media platforms aiming to balance free speech with healthy discourse.



This visualization highlights smaller, tightly-knit clusters, showing us exactly how users interact within niche conversations. Identifying these structural divisions helps platforms recognize isolated groups where toxicity can grow unnoticed, improving targeted moderation strategies. Girvan-Newman specifically looks for weak points or bottlenecks in communication [1], places

where the removal of certain connections would break the network into smaller segments. The visualization shows that there are many small groups represented by distinct colors, mostly concentrated on the network's outer edges.

The center of the network, dominated by purple-colored nodes, represents larger groups of highly interconnected users, suggesting these central communities actively drive most conversations. This contrasts with Louvain community detection, which groups users into broader communities based on overall interaction patterns, rather than highlighting specific bottlenecks. Thus, while Louvain gives a broader structural perspective, the Girvan-Newman method helps pinpoint critical interaction hubs and reveals more detailed subgroup dynamics, offering valuable clues on managing and understanding community behaviors, including the spread and containment of toxic interactions.

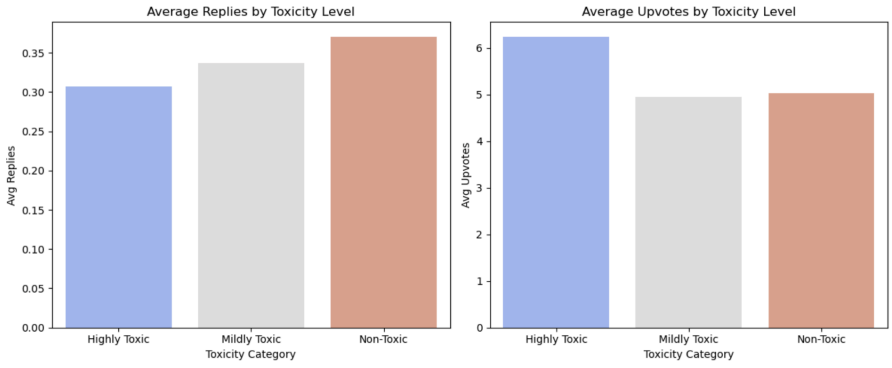
Community ID	Avg Toxicity	Proportion of Highly Toxic Users
0	0.229104	0.225
1	0.001643	0.000
2	0.944809	1.000
3	0.000593	0.000
4	0.441162	0.452

In this part, I analyzed the Reddit interaction network data to explore patterns and insights about user behavior, toxicity, and community structures. After constructing the user interaction network, I applied several network analysis techniques to examine how toxicity impacts the conversation flow. Specifically, I used toxicity scores (obtained from the NLP model "Detoxify") to classify communities beyond traditional structural methods. By shifting our focus from traditional structural community detection (like Louvain and Girvan-Newman methods, which group users based on their interactions alone) to a more behavioral-focused classification, I reorganized the user communities into three clear categories—Non-Toxic, Mildly Toxic, and Highly Toxic. This allowed a deeper understanding of how toxic behavior influences community engagement and network dynamics.

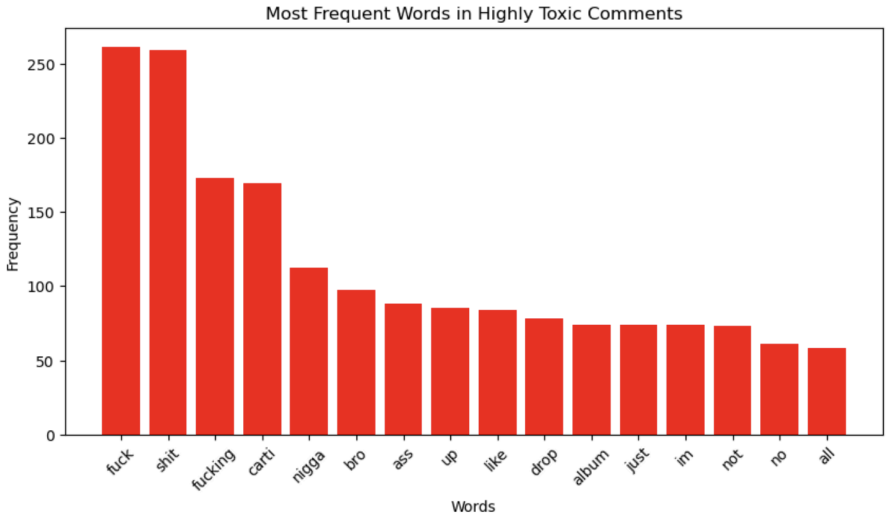
The full network statistics give us an overall snapshot of how Reddit users interact with each other. We have 1,792 users (nodes) who made 1,756 interactions (edges). The average number of connections per user is around 1.96, which means most users interact with only about two others, on average. The network density (0.000547) is extremely low, telling us that most possible connections between users don’t exist, and users generally talk to only a small set of other people. Also, the graph isn't fully connected, meaning there are groups or users who never interact with others outside their own circle. The average clustering coefficient (0.0133) is low, showing that Reddit conversations mostly happen between pairs of users, not within tightly-knit groups. These findings are helpful for understanding our toxicity analysis. Because the network isn't tightly connected, toxic behavior probably stays isolated within small groups

rather than spreading widely across the network. This means we likely have small pockets of toxicity rather than a large-scale issue affecting everyone.

Furthermore, centrality analysis helped me pinpoint influential users in the network. Users such as "TinyCatIsABoss" and "NoPanic3036" had the highest degree of centrality, suggesting they interacted frequently with many others and played active roles in conversations. Users like "RnRClub44" and "CatchASvech," who scored high in betweenness centrality, acted as bridges connecting different sub-communities, significantly influencing information flow. Eigenvector centrality analysis identified influential users like "EvieGoesHard" and "NoPanic3036," highlighting individuals whose interactions were with other well-connected users, thus amplifying their influence within the network.



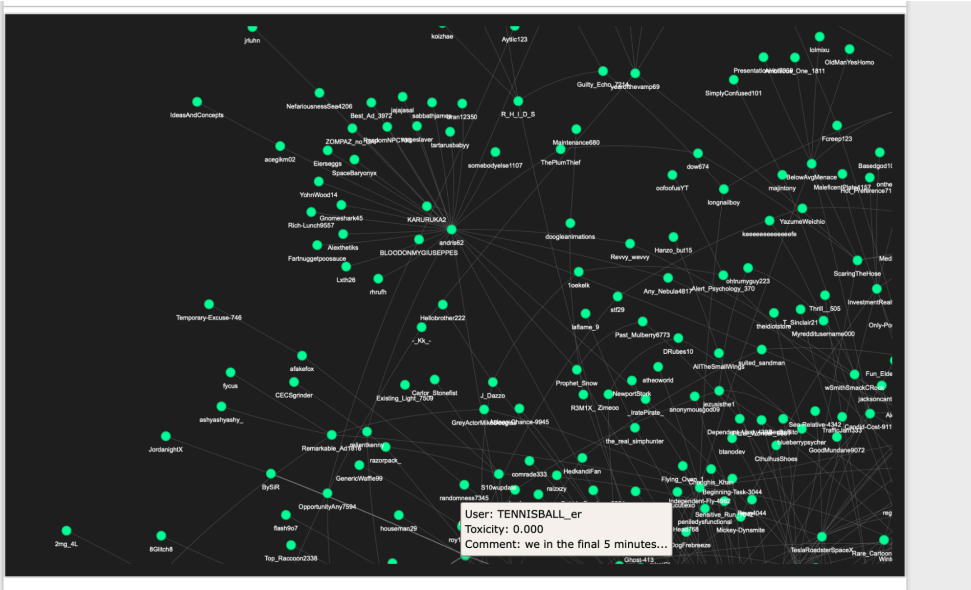
Visualization-1: Replies vs Toxicity and Upvotes vs Toxicity.



Visualization-2

Finally, examining the engagement metrics (upvotes and replies) provided crucial insights into user behavior patterns. Highly toxic comments received fewer replies but noticeably higher average upvotes. This suggests that while people might avoid openly interacting with negativity, they still silently endorse or amplify toxic comments through upvotes. Furthermore, an analysis

of common toxic words indicated specific problematic language frequently used to provoke or intensify discussions. These findings have clear practical implications, as they provide guidance for improving moderation practices and designing healthier online community environments by proactively managing toxicity spread.



The interactive visualization (provided html version in github) presents an engaging and dynamic view of the Reddit user interaction network, colored based on toxicity levels. Users are represented as nodes, and their interactions as edges, with colors reflecting varying degrees of comment toxicity. Darker colors typically indicate higher toxicity, allowing us to intuitively identify and analyze toxic hubs or influential users within the network. This visualization enables an interactive exploration, giving insights into specific user interactions, comments, and their associated toxicity, thus effectively merging structural and behavioral analyses in a visually appealing and accessible manner.

From this assignment, I realized how toxicity actually affects discussions on Reddit. Toxicity isn't just stuck in one place—it moves across different groups, affecting how people interact. Interestingly, toxic comments grab attention easily because they get lots of upvotes, but people don’t usually reply to them much, maybe because they don't want direct conflict. By looking at the network structure, I saw that toxic users tend to stay on the edges rather than in the middle, indicating that they're less actively involved but still noticeable. Through this analysis, I also identified influential users who might unintentionally spread toxicity because they're well-connected. Understanding these patterns is valuable—it can help social media platforms moderate content better, encourage healthier conversations, and prevent toxic behavior from spreading further.

References:

1. M. Girvan, & M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. U.S.A. 99 (12) 7821-7826, https://doi.org/10.1073/pnas.122653799 (2002).

Tools:

Generative AI.